心理学报 2022, Vol. 54, No. 9, 1137−1150
Acta Psychologica Sinica

https://doi.org/10.3724/SP.J.1041.2022.01137

考虑题目选项信息的非参数认知诊断计算机自适应测验*

孙小坚[1,2,3] 郭 磊[3,4]

([1]西南大学数学与统计学院; [2]西南大学基础教育研究中心;
[3]中国基础教育质量监测协同创新中心西南大学分中心; [4]西南大学心理学部, 重庆 400715)

摘 要 选择题中的作答选项能提供额外诊断信息, 为充分利用选项信息, 研究提出认知诊断计算机自适应测验(CD-CAT)中两种处理选择题选项信息的非参数选题策略和变长终止规则。模拟研究的结果发现：(1)定长条件下两种非参数选题策略的分类准确性整体要高于参数选题策略; (2)两种非参数选题策略较参数选题策略具有更加均衡的题库使用情况; (3)非参数选题策略在两种新的变长终止规则下具有更高的分类准确率; (4)两种非参数选题策略均适用于选择题 CD-CAT 情境, 使用者可任选其一进行测验分析。

关键词 认知诊断计算机自适应测验, 题目选项信息, 非参数选题策略, 变长终止规则
分类号 B841

1 引言

认知诊断评估(cognitive diagnostic assessment, CDA)是对个体的知识、技能以及认知加工过程进行诊断分类的一种方法。其可以提供个体在各知识内容上的具体掌握情况, 故可知晓个体学习过程中的优势与不足(辛涛 等, 2015), 这一方面有利于解释个体在某些测验上表现欠佳的原因, 同时还有利于教师进行后续的补救教学(如 Gao et al., 2021), 因此受到众多研究者的重视。CDA 中对个体进行分类的方法主要有参数方法和非参数方法(郭磊, 周文杰, 2021), 参数方法主要使用认知诊断模型(cognitive diagnostic model, CDM)估计题目参数和个体属性掌握情况, 其中, 一般性 CDM 有广义 DINA 模型(generalized deterministic inputs, noisy “and” gate, GDINA; de la Torre, 2011)等, 简化模型则有 DINA 模型等。非参数方法主要有聚类分析法(郭磊 等, 2018; 康春花 等, 2015; Chiu et al., 2009)、距离判别法(康春花 等, 2019; 罗照盛 等, 2015; Chiu et al., 2018)以及机器学习法(汪文义 等, 2016; Liu & Cheng, 2018)。

当前 CDA 的一个重要研究领域是认知诊断计算机化自适应测验(cognitive diagnostic computerized adaptive testing, CD-CAT; Chang, 2015, Cheng, 2009)。相较于纸笔测验, CD-CAT 能够以更少的题目获得更加准确的诊断结果, 并且其提供的题目测量特性与个体知识掌握水平大体相当, 因而可较好地激发个体的作答动机(陈平 等, 2011; 孙小坚 等, 2019; Sun et al., 2021), 最终实现对个体的准确测量。CD-CAT 包含 5 个重要组成部分：题库、测验模型、选题策略、知识状态估计方法以及终止规则。其中选题策略受到大量关注(郭磊 等, 2016); 常见的选题策略有基于作答分布和基于后验分布的策略(Zheng & Chang, 2016), 前者包括 KL 信息(Kullback-Leibler information, Xu et al., 2003)、GDI (G-DINA model discrimination index, Kaplan et al., 2015)等;

收稿日期: 2021-12-31
* 国家自然科学基金青年项目(31900793)、中央高校基本科研业务费专项资金(SWU2109222)、中国科普研究所委托项目(210105ESR056)、2020 年度重庆市博士后特别资助项目资助。
通信作者: 郭磊, E-mail: happygl1229@swu.edu.cn

后者包括 SHE (Shannon entropy, Xu et al., 2003)等。

以上选题策略属于参数策略，参数选题策略的优势在于可以准确获得个体在各属性掌握模式(attribute master pattern, AMP)上的掌握概率，同时也可以获得较高的分类准确性。而其潜在不足在于由于需知晓题库中的题目参数，要求事先进行预测试，如此，若样本量不够大，则题目参数的精度难以保障，同时还存在题库被提前泄漏的风险。对此，研究者提出了非参数 CD-CAT (何明霜, 2021; 张淑君, 2019; Chang et al., 2019; Chiu & Chang, 2021)。非参数 CD-CAT 只需获得题库中各测验题目所考察的属性，无需进行预测试，因而降低了题库提前泄漏的风险，且无需考虑题目参数估计偏差带来的影响(Chang et al., 2019)。模拟研究的结果表明，当预测试样本较小时，非参数选题策略下的模式匹配率(pattern matched rate, PMR)要优于参数选题策略的 PMR (Chang et al., 2019; Chiu & Chang, 2021; Yang et al., 2020)。

终止规则作为 CD-CAT 的另外一个重要成分，受关注程度远低于选题策略，且大部分 CD-CAT 研究使用定长终止规则，对变长 CD-CAT 的研究较少(郭磊 等, 2015)。但相较于定长 CD-CAT，变长 CD-CAT 在测验效率、能力估计的收敛情况和估计精度等方面均更加优异，更能体现自适应测验的特点和优势(郭磊 等, 2015)。对此，Hsu 等(2013)基于具有最大和次大后验概率的两个 AMPs 之间的比值提出一种变长终止规则；郭磊等(2015)提出了 6 种变长终止规则，其进一步将此 6 种规则分成基于绝对标准、基于相对标准以及结合绝对和相对标准的混合终止规则；上述终止规则适用于参数 CD-CAT，难以拓展至非参数 CD-CAT 情境。张淑君(2019)则在非参数 CD-CAT 情境下提出两种变长终止规则(D1 和 D3)，并通过模拟和实证研究对两种规则的表现进行了探究，结果表明 D3 规则的 PMR 要明显高于 D1 规则的 PMR。值得注意的是，一方面，当前关于非参数终止规则的研究依旧非常少；另一方面，张淑君(2019)提出的 D3 终止规则思路在于每次估计个体的 AMP 时，最小非参数距离(如汉明距离)只能对应一个 AMP，但 3 次估计的 AMPs 之间可能各不相同，如此可能增加估计误差。

当前无论是参数还是非参数 CD-CAT，绝大多数研究的主流题型是选择题(multiple-choice, MC)，且对个体的作答反应进行分析时重点关注个体是否正确作答测验题目，而较少关注干扰项信息，忽视了干扰项所能提供的额外诊断信息，导致对 MC 题目的使用效率过低(郭磊, 周文杰, 2021; 刘拓, 2016; de la Torre, 2009)。对此，Yigit 等(2019)基于选择题 DINA 模型(MC-DINA; de la Torre, 2009)提出了同时考虑所有选项的 JSD (Jensen–Shannon divergence)选题策略，研究结果表明 JSD 策略下的 PMR 较不考虑干扰项信息的 GDI 具有更高的 PMR。但 JSD 选题策略属于参数策略，意味着题目参数估计偏差和题库泄露风险问题依然存在。考虑到非参数诊断方法无需或只需少量预测试样本即可获得较高的 PMR (Chang et al., 2019; Chiu & Chang, 2021)，且当前尚未有研究探讨非参数方法如何在 CD-CAT 中利用干扰项信息以提升对个体的诊断精度。基于此，本研究一方面提出两种 CD-CAT 中融合干扰项信息的非参数选题策略，同时，为更好地实现 CD-CAT 的自适应特点，提出两种适用于考虑题目选项信息的 CD-CAT (记为 mcCD-CAT)的非参数变长终止规则。研究将通过模拟研究分别对二种非参数选题策略和变长终止规则的性能进行系统探讨，以进一步丰富 CD-CAT 研究。文章的结构如下：首先介绍可处理选项信息的认知诊断方法，其次阐述非参数 mcCD-CAT 及其变长终止规则，之后通过两个模拟研究探讨非参数 mcCD-CAT 和终止规则的性能，最后对结果进行讨论与展望。

2 处理选项信息的认知诊断方法

目前研究者提出了参数和非参数的认知诊断方法以处理考虑题目选项信息的认知诊断测验，下面对此二者进行介绍。

2.1 MC-DINA 模型

研究者提出了选择题 CDMs 以处理选项信息，如 MC-DINA 模型、SICM 模型(scaling individuals and classifying misconceptions model, Bradshaw & Templin, 2014)和结构化 MC-DINA 模型(Ozaki, 2015)等。其中 SICM 模型将个体的潜在特质看作是连续变量，而各干扰项则是关于知识内容的错误概念，这与常规的 CDA 存在差异，故研究不考虑该模型。考虑到 MC-DINA 模型简单易懂，参数的解释性也更加通俗易懂，且具有不错的诊断效果，故研究将介绍该模型。MC-DINA 模型的作答反应函数为：

$$P_{jh}(\boldsymbol{\alpha}_i) = P(X_{ij} = h \mid \boldsymbol{\alpha}_i) = P(X_{ij} = h \mid g_{ij} = g) \triangleq P_j(h \mid g) \tag{1}$$

$$g_{ij} = \arg\max_h \left\{ \boldsymbol{\alpha}_i^T \boldsymbol{q}_{jh} \mid \boldsymbol{\alpha}_i^T \boldsymbol{q}_{jh} = \boldsymbol{q}_{jh}^T \boldsymbol{q}_{jh} \right\}$$

其中, $\boldsymbol{\alpha}_i$ 表示个体 i 的 AMP; j 表示题目, h 表示选项, $h=1,\cdots,H_j$, H_j 表示题目 j 的选项个数; g_{ij} 表示个体 i 在题目 j 上所属的组别, $\boldsymbol{q}_{jh}$ 表示题目 j 中第 h 个选项的 q 向量, $\boldsymbol{\alpha}_i^T\boldsymbol{q}_{jh}$ 表示 $\boldsymbol{\alpha}_i$ 和 $\boldsymbol{q}_{jh}$ 的内积, 上标 T 表示转置; $g_{ij}=0,1,\cdots,H_j^*$, 其中, H_j^* 表示题目 j 中选项元素不全为 0 的选项个数, 当个体 i 的 AMP 与题目 j 的所有选项均无法匹配时, $g_{ij}=0$。$P_j(h|g)$ 则表示属于组别 g 中的个体在题目 j 上选择第 h 个选项的概率。为保证模型可被识别, 通常限定 $\sum_{h=1}^{H_j}P_j(h|g)=1$, 即组别 g 中的个体在题目 j 上选择各选项的概率总和等于 1。不同于 DINA 模型, MC-DINA 模型的参数是选项的选择概率 $P_j(h|g)$ 本身。

通过相应的参数估计方法如 EM 算法(de la Torre, 2009)、MCMC 算法(Ozaki, 2015)和 VB 算法(Variational Bayesian; Yamaguchi, 2020)即可得到 MC-DINA 模型的参数和个体的 AMP 估计值。

2.2　MC 汉明距离法

郭磊和周文杰(2021)提出了基于选项层面的非参数诊断分类方法, 其使用汉明距离计算观测作答和题目各选项的理想作答间的距离总和, 距离总和最小的 AMP 将作为个体的 AMP。在此基础上, 他们提出了简单 MC 汉明距离、加权 MC 汉明距离以及惩罚 MC 汉明距离三种非参数方法, 其中后两者是简单 MC 汉明距离的拓展。其模拟研究发现, 简单 MC 汉明距离下的 PMR 要优于另外二者, 故此处将介绍简单 MC 汉明距离, 其表达式为:

$$HDDmc(\mathbf{X}_i,\boldsymbol{\eta}_i)=\sum_{j=1}^{J}\sum_{h=1}^{H_j}\left|X_{ijh}-\eta_{ijh}\right| \tag{2}$$

$$\eta_{ijh}=\left[\prod_{k=1}^{K_j^*}2-2^{(\alpha_{ik}-q_{jhk})^2}\right]\times\left[1-\prod_{k=1}^{K_j^*}(1-\alpha_{ik})\right]$$

其中, X_{ijh} 和 η_{ijh} 分别表示个体 i 在题目 j 第 h 个选项上的实际作答和理想作答, 二者取值均为 0 或 1, 表示个体是否选择该选项; K_j^* 表示题目 j 所考察的属性个数。

3　考虑题目选项信息的 CD-CAT

目前关于 mcCD-CAT 的研究比较少, Yigit 等(2019)在 MC-DINA 模型的基础上提出了参数 JSD 策略。研究首先对 JSD 策略进行介绍, 然后再介绍本研究提出的两种适用于 mcCD-CAT 的非参数策略。非参数 CD-CAT 将基于作答反应与各 AMPs 之间的非参数距离(如汉明距离)对个体进行分类并且选择后续的测验题目(Chang et al, 2019; Chiu & Chang, 2021), 因而计算作答反应和所有 AMPs 间的距离是非参数 CD-CAT (包括 mcCD-CAT)的核心和基础。

3.1　基于 JSD 的 mcCD-CAT

Yigit 等(2019)以 MC-DINA 模型为基础, 提出了可以考虑所有选项信息的 JSD 策略, JSD 策略是一种基于作答反应后验分布的选题策略, 通过相应的转换, 其与 SHE 策略等价。JSD 策略的计算公式为:

$$JSD_j=S(\mathbf{P}_j\times\boldsymbol{\pi}^T)-\sum_l\pi_l S(\mathbf{P}_{jl})=$$
$$S\left[\sum_l P(X_j=h|\boldsymbol{\alpha}_l)\pi(\boldsymbol{\alpha}_l)\right]-\sum_l\pi(\boldsymbol{\alpha}_l)S(\mathbf{P}_{jl})=$$
$$-\sum_h P(X_j=h)\log P(X_j=h)-$$
$$\sum_l\pi(\boldsymbol{\alpha}_l)S(P(X_j=h|\boldsymbol{\alpha}_l)$$

其中 $\mathbf{P}_j$ 为 $H_j\times 2^K$ 的概率矩阵, 表示所有 AMPs 选择各个选项的概率; $\mathbf{P}_{jl}$ 为 $H_j\times 1$ 的向量, 表示第 l 种 AMP 在题目 j 上选择各个选项的概率; $\boldsymbol{\pi}$ 为 $2^K\times 1$ 的向量, 表示各 AMP 的后验概率; S(·)表示香农熵: $S(x)=E[-\log x]$; $P(X_j=h)$ 表示个体选择题目 j 第 h 个选项的边际概率。候选题目集中具有最大 JSD 值的题目将提供给个体作答。

3.2　基于 MC 汉明距离的 mcCD-CAT

郭磊和周文杰(2021)提出的 η_{ijh} 是对 Ozaki (2015)文章所提指标的修正, 该值的计算过程要求个体 i 对题目 j 所考察属性的掌握情况与选项 h 的缩减 q 向量(即 $\underbrace{\{1,\cdots,1\}}_{K_j^*}$, K_j^* 为正确选项所考察的属性个数)完全匹配, 其计算过于严苛。例如, 假设题目 j 考察 3 个属性, 其在 4 个选项上的缩减 q 向量分别为{1, 1, 1}、{1, 0, 1}、{1, 0, 0}和{0, 0, 0}, 此时, 基于郭磊和周文杰(2021)的 η_{ijh} 计算方法, 缩减 AMPs 分别为{1, 1, 0}和{1, 0, 0}的个体在该题目上的理想作答向量分别为{0, 0, 0, 1}和{0, 0, 1, 0}。但理论上, AMP 为{1, 1, 0}的个体由于未掌握第 3 个属性, 所以其选择第 1 和 2 个选项的概率较小, 而第 4 个选项未考察任何属性, 故该个体也不大可能选择该选项, 第 3 个选项考察了第一个属性, 而该个体也掌握了第一个属性, 故理想状态下, 第 3 个选项是其理想作答(de la Torre, 2009)。对此, 本研究将对上述 η_{ijh} 的计算方法进行完善, 使其更好地

适应测验情境。由于个体所属组别 g_{ij} 和 H_j^* 之间存在对应关系，而 H_j^* 是对不同选项的 q_{jh} 向量的表征，由此可得 g_{ij} 也是对 q_{jh} 的表征，而 η_{ijh} 同样是对 q_{jh} 的表征，因此 g_{ij} 和 η_{ijh} 之间有内在关系。此时，将 η_{ijh} 定义为：

$$\eta_{ijh}=\begin{cases}1 & 如果 g_{ij}=h\\ 0 & 其它\end{cases} \tag{3}$$

其中，g_{ij} 表示个体所属的组别，其值可由公式(1)计算得到。

基于修改的 η_{ijh} 重新考虑上面的例子：先用公式(1)计算缩减 AMPs 分别为{1, 1, 0}和{1, 0, 0}的个体在题目 j 中的组别，此时，二者的 g_{ij} 均为 3；再基于公式(3)，得到二者的 η_{ijh} 均为{0, 0, 1, 0}；最后代入公式(2)计算其 HDDmc 值。

基于 MC 汉明距离的 mcCD-CAT 实施流程为：(1)初始化测验题库，明确各测验题目及选项所考察的属性；(2)从题库中随机选择一个测验题目给个体作答，记录个体的作答反应；(3)计算个体当前作答反应向量和所有 AMPs 间的距离(HDDmc)；(4)基于 HDDmc 值对所有 AMPs 进行升序排序，并确定顺序为前两位的 AMPs (分别记为 $\hat{\boldsymbol{\alpha}}_{1st}$ 和 $\hat{\boldsymbol{\alpha}}_{2nd}$)，其对应的距离分别记为 d_{1st} 和 d_{2nd}；(5)从测验题库中筛选出能够区分 $\hat{\boldsymbol{\alpha}}_{1st}$ 和 $\hat{\boldsymbol{\alpha}}_{2nd}$ 的题目集 S，即 S 中的题目满足 $\boldsymbol{\eta}_{j,\hat{\alpha}_{1st}}\neq\boldsymbol{\eta}_{j,\hat{\alpha}_{2nd}}$，若题目集 S 为非空集合，则转至步骤(7)，反之则转至步骤(6)；(6)用 $\hat{\boldsymbol{\alpha}}_{m^{th}}$ ($m \geqslant 3$)替换 $\hat{\boldsymbol{\alpha}}_{2nd}$，并筛选出候选题目集合 S；当 S 为空集时，用 $m = m + 1$ 更新 m 值，并重复步骤(6)，直到 S 为非空集合；(7)从题目集 S 中随机抽取一个题目给个体作答，记录其作答反应；(8)重复步骤(3)到(7)，直到满足终止规则；(9)将 d_{1st} 所对应的 AMP 作为个体最终的估计值。

由以上流程可知，mcCD-CAT 实施流程和 Chang 等(2019)提出的非参数 CD-CAT 实施流程大体相同，不同之处在于 Chang 等(2019)使用了 Xu 等(2016)提出的初始题选择策略，而 mcCD-CAT 则无需该步骤，其原因在于 mcCD-CAT 将题目所有选项纳入考虑，而各选项中的 q 向量可以较好地体现 Xu 等(2016)中的初始题选择策略，因此，在测验长度较长的情况下，mcCD-CAT 自身已经蕴含了 Xu 等(2016)的初始题选择策略，故不需要重复进行初始题的选择，预实验研究和后面的模拟研究 1 均证明了这点。

3.3 基于 Jaccard 距离的 mcCD-CAT

Jaccard 相似度(Jaccard similarity; Jaccard, 1912)最初应用于植物学领域，用于测量两个不同区域 A 和 B 的植物种类间的相似程度，后被广泛应用于信息检索、数据挖掘和机器学习等领域(Kosub, 2019)；何明霜(2021)将其应用于多级计分的 CD-CAT，本研究将其拓展至 mcCD-CAT 情境。Jaccard 相似度的计算方法为(Jaccard, 1912)：

$$Jac=\frac{|A\cap B|}{|A\cup B|}=\frac{n_{AB}}{n_A+n_B+n_{AB}}$$

其中，n_A 和 n_B 分别表示区域 A 和 B 中独有的物种数量，而 n_{AB} 则表示两个区域共有的物种数量。Jac 取值范围为[0, 1]，0 和 1 分别表示完全不一致和完全一致。本研究将其用于计算观察作答反应和理想作答反应之间的相似度，并基于相似度值来对个体进行诊断分类，为使 Jac 值与 HDDmc 有相同形式，研究使用 $1 - Jac$ 表示相似度(也称 Jaccard 距离)：

$$JDDmc(\boldsymbol{X}_i,\boldsymbol{\eta}_i)=1-\frac{|\boldsymbol{X}_i\cap\boldsymbol{\eta}_i|}{|\boldsymbol{X}_i\cup\boldsymbol{\eta}_i|}=1-\frac{\sum_{j=1}^{J}\sum_{h=1}^{H_j}I(X_{ijh}=\eta_{ijh})}{\sum_{j=1}^{J}H_j} \tag{4}$$

其中，$\boldsymbol{X}_i=\{\boldsymbol{X}_{i1},\cdots\boldsymbol{X}_{ij},\cdots,\boldsymbol{X}_{iJ}\}$ 和 $\boldsymbol{\eta}_i=\{\boldsymbol{\eta}_{i1},\cdots\boldsymbol{\eta}_{ij},\cdots,\boldsymbol{\eta}_{iJ}\}$ 分别表示个体 i 的实际和理想作答反应模式，$\boldsymbol{X}_{ij}$ ($\boldsymbol{\eta}_{ij}$)表示个体 i 在题目 j 上的实际(理想)作答模式，是长度为 H_j 的二分向量，如 $\boldsymbol{X}_{ij}=\{0,1,0,0\}$ 表示个体选择了第 2 个选项。J 表示个体作答的题目数量，$I(\cdot)$为指示函数，表示括号内的表达式是否成立，成立为 1，反之为 0。文中 JDDmc 的计算过程与何明霜(2021)的计算过程之间的主要差异在于理想作答模式的计算，本文中的理想作答模式的计算见公式(3)。

需注意的是，由于事先并不清楚个体的 AMP，故无法直接获得 $\boldsymbol{\eta}_l$，此时，将依次计算所有可能的 AMPs 在这些题目上的理想反应 $\boldsymbol{\eta}_l$，$l=1,\cdots,2^K$，并计算 $\boldsymbol{\eta}_l$ 和 $\boldsymbol{X}_i$ 之间的 JDD 值，个体最终的 AMP 具有最小的 JDD 值，若最小 JDD 值对应多个 AMPs，则从中随机选择一个。

基于 JDDmc 的 mcCD-CAT 实施流程和基于 HDDmc 的 mcCD-CAT 实施流程基本相同，不同之处在于步骤(3)，此时使用 Jaccard 距离计算公式来计算个体实际作答反应向量和所有 AMPs 间的非参数距离。

3.4 终止规则

CD-CAT 的终止规则分为定长和变长两类。当测验为定长时，其终止规则为预先设定的题目长度，这在非参数和参数 CD-CAT 中均适用；当测验为变

长时，张淑君(2019)在非参数 CD-CAT 中提出 D1 和 D3 两种终止规则，其思路是每次估计个体 AMP 时，最小距离(如 HDDmc)是否对应唯一的 AMP。D1 规则下，个体作答某题目后，当最小 HDDmc 只对应一个 AMP 时，结束测验；D3 规则下，每次估计个体AMP时，要求具有最小HDDmc只对应唯一的AMP，且这种一一对应关系需连续出现3次才能结束测验。

本研究基于限制性 MHRM 算法(cMHRM；Liu et al., 2020)和基于距离比的思路提出两种适用于非参数 CD-CAT 的变长终止规则(分别记为 MR 和 DR 规则)，以丰富此方面的研究。Liu 等(2020)使用 cMHRM 算法估计 CDM，该算法需计算前后两次迭代的所有参数估计值间的差值 $\boldsymbol{\delta}$，并取最大差值 *max* (δ)，将每次迭代的 *max* (δ)组成向量 $\Delta=\{\max(\delta^1),\cdots,\max(\delta^t)\}$，当 $\boldsymbol{\Delta}$ 中连续 4 个 *max*(δ)均小于预设标准时，算法结束。本研究将借鉴该思想：当连续 4 次所估计的 AMPs 均相同时，测验终止，并将该 AMP 作为个体最终的 AMP。第二种变长终止规则是计算 d_{1st} 和 d_{2nd} 之间的比值，该方法的思想在因子分析中抽取单个因子时经常被使用。本研究通过计算 $d_{2nd}/d_{1st}>CR$ (CR 为预设值)来终止测验，并记 d_{1st} 所对应的 AMP 为个体最终的 AMP。

4　研究 1：定长 mcCD-CAT 下两种非参选题策略的性能

4.1　研究目的

在固定测验长度条件下，探讨两种考虑干扰项信息的非参数选题策略在不同实验条件中的性能，并将其与参数选题策略(JSD)进行比较。

4.2　研究设计

4.2.1　自变量

研究的自变量有 6 个，分别为属性个数、Q 矩阵结构、题目质量、属性分布形态、测验长度和选题策略。具体而言，(1)属性个数分别为 4 和 6 个，4 和 6 个属性在以往研究中比较常见(如：孙小坚 等 2019, 2021; Sun et al., 2021)。(2) Q 矩阵的结构有两种，分别为简单结构和复杂结构(郭磊 等，2015)，其中简单结构下，题目的正确选项考察各属性的概率为 20%，且正确选项至少考察一个属性；复杂结构下，题目正确选项考察各属性的概率则为 50%。错误选项的 q 向量则为正确选项的子集，且选项之间具有包含关系(de la Torre, 2009)。(3)题目质量有 3 个水平，分别为高、低和混合质量，题目质量将通过 $1-P_j(h\mid g)$ 给予表征，3 种质量分别服从以下均匀分布(Sun et al., 2020)：*U* (0.05, 0.25)、*U* (0.25, 0.45)和 *U* (0.05, 0.45)，剩余选项平均分配 $1-P_j(h\mid g)$ 值，以保证 $\sum_{h=1}^{H_j}P_j(h\mid g)=1$。(4)属性分布形态有两种，分别为多元正态阈值模型和均匀分布(如：郭磊，周文杰，2021；Chang et al., 2019; Chiu & Chang, 2021)。(5)测验长度有 3 个水平，由于涉及不同属性个数，故研究针对属性个数进行测验长度的设定，3 种测验长度分别为 2K、3K 和 4K，其中 K 表示属性个数。(6)选题策略有 3 个水平，分别为 HDDmc、JDDmc 和 JSD。

4.2.2　控制变量

研究的控制变量主要有测验模型、题库大小、选项数量、正式测试的人数。研究将用 MC-DINA 模型生成作答数据(Yigit et al., 2019)，选择该模型的原因在于，首先，可处理题目选项信息的饱和 CDM 非常少，相关研究也不成熟，且参数难以解释和估计；其次，当前绝大多数 CD-CAT 的研究采用简化模型如(DINA)进行分析，只有极少量研究使用饱和模型；最后，相关的实证研究亦采用 DINA 模型进行 CD-CAT 分析(如 Liu et al., 2013)。题库方面则固定题库中的题目数量为 480 (孙小坚 等，2021)。选项个数固定为 4 个，这在实际测验中较为常见。正式测试的人数则固定为 500 人(Chang et al., 2019)。此外，参考以往研究(如 Chang et al., 2019; Chiu & Chang, 2021; Yang et al., 2020)，使用 JSD 时，先基于预测试进行参数校准，此时校准的样本量固定为 40K，其中 K 为属性个数；校准完毕后，将基于校准的题目参数选择最佳的侯选题目。

研究总共有 2 (属性个数) × 2 (Q 矩阵结构) × 3 (题目质量) × 2 (属性分布形态) × 3 (测验长度) × 3 (选题策略) = 216 种实验条件，其中选题策略为被试内变量，其它则为被试间变量。为减少抽样误差，各实验条件重复 30 次。所有程序用 R 软件实现。

4.3　评价指标

评价指标有两类，一类用于评价诊断分类的准确性，用 PMR 体现，其值在 0 和 1 之间，值越大则分类越准确；另一类则用于评价题库使用情况，包括测验整体曝光率 χ^2，测验重叠率(TOR)、曝光不足率(UIR)和过度曝光率(OIR)，四者越小越好(陈平 等，2011；孙小坚 等，2021)。以上指标的计算公式为：

$$PMR=\frac{1}{R}\sum_{r=1}^{R}\left[\sum_{i=1}^{N}I(\hat{\boldsymbol{\alpha}}_i=\boldsymbol{\alpha}_i)\Big/N\right]$$

$$\chi^2 = \frac{1}{R}\sum_{r=1}^{R}\left[\sum_{j=1}^{N_{item}}(\exp_j - J/N_{item})^2 \Big/ (J/N_{item})\right]$$

$$\exp_j = N_j^a / N$$

$$TOR = \frac{1}{R}\frac{\sum_{r=1}^{R}\left[\sum_{j=1}^{N_{item}} N_j^a \times (N_j^a - 1)\right]}{J \times N \times (N-1)}$$

$$UIR = \frac{1}{R}\sum_{r=1}^{R}\left[\sum I(\exp_j < 0.02) \Big/ N_{item}\right]$$

$$OIR = \frac{1}{R}\sum_{r=1}^{R}\left[\sum I(\exp_j > 0.2) \Big/ N_{item}\right]$$

其中，R 为重复次数，$\hat{\boldsymbol{\alpha}}_i$ 和 $\boldsymbol{\alpha}_i$ 分别表示估计和真实的 AMP；N_j^a 则为题目 j 被使用的次数。

4.4 研究结果

4.4.1 HDDmc 和 JDDmc 的分类准确性整体优于 JSD

图 1 呈现了 4 个属性下 3 种选题策略在不同实验条件下的 PMRs。整体而言，两种非参数策略(HDDmc 和 JDDmc)的估计准确性在所有条件下基本相同，并且二者在绝大多数条件下的 PMRs 要高于 JSD 方法。具体而言，在题目质量为高和混合条件下，HDDmc 和 JDDmc 的 PMRs 整体要高于 JSD 方法，并且随着测验长度的增加，HDDmc 和 JDDmc 与 JSD 间的 PMR 差异不断增大。在简单 Q 矩阵和低题目质量条件下，JSD 与 HDDmc 和 JDDmc 之间的差异比较小，在部分条件下 JSD 的 PMR 略微高于 HDDmc 和 JDDmc；但在复杂 Q 矩阵条件下，HDDmc 和 JDDmc 的 PMR 要明显高于 JSD 方法，只在两个条件(混合题目质量下测验长度为 2K 和 3K)下的 PMR 与 JSD 相同或相近。此外，题目质量和测验长度对 3 种选题策略具有积极影响，题目质量越高、测验长度越长，则 3 种策略的 PMR 越高。另外，非参数方法在复杂 Q 矩阵下的 PMRs 整体高于简单 Q 矩阵的结果。

6 个属性下 3 种选题策略在不同实验条件下的 PMRs 如图 2 所示。简单 Q 矩阵条件下，HDDmc 和 JDDmc 在 3 个条件下的 PMRs 高于 JSD，而在剩余 6 个条件下的 PMRs 则低于 JSD，特别是在混合题目质量下，二者与 JSD 在 PMRs 上存在比较大的差异。复杂 Q 矩阵条件下，HDDmc 和 JDDmc 的 PMRs 则在大多数条件下高于 JSD 策略，只在混合题目质量和 2K 个题目长度下的 PMR 小于 JSD。当题目质

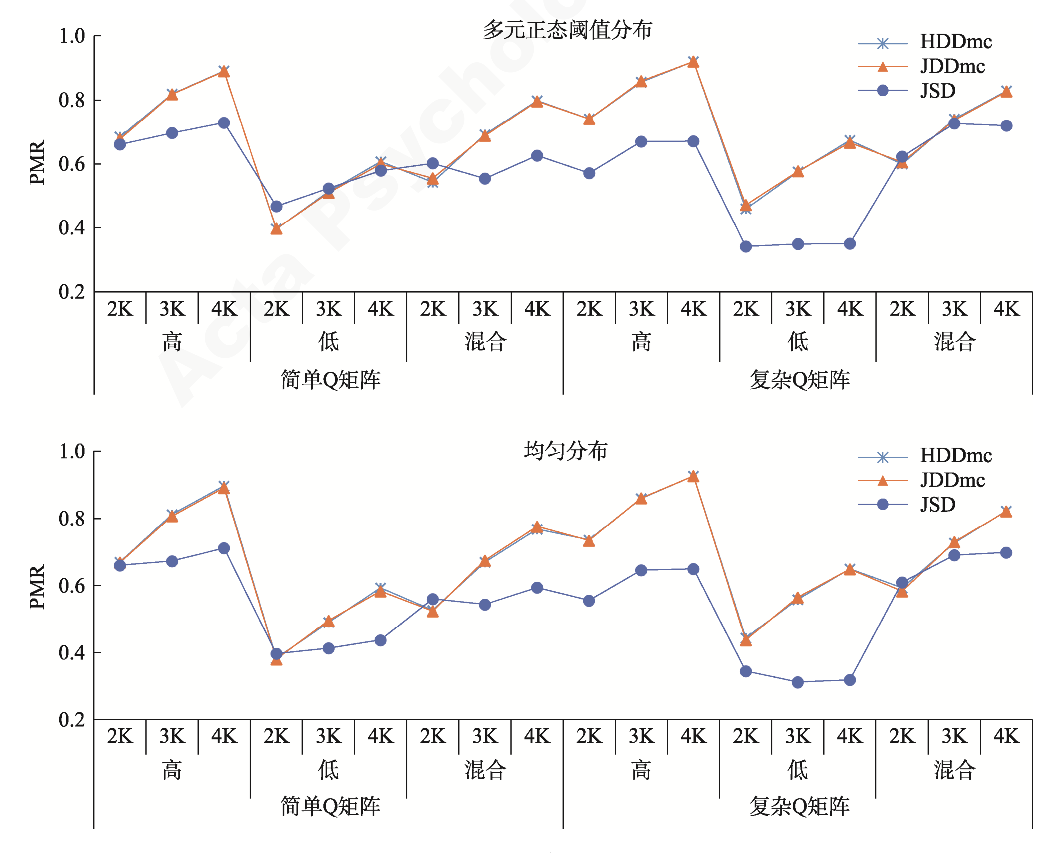


图 1 四个属性下各条件的分类准确性

注：2-4K 表示测验长度为属性个数的 2-4 倍；HDDmc 为基于 MC 汉明距离的选题策略，JDDmc 为基于 Jaccard 距离的选题策略，JSD 为 基于 JSD 的选题策略，下同。

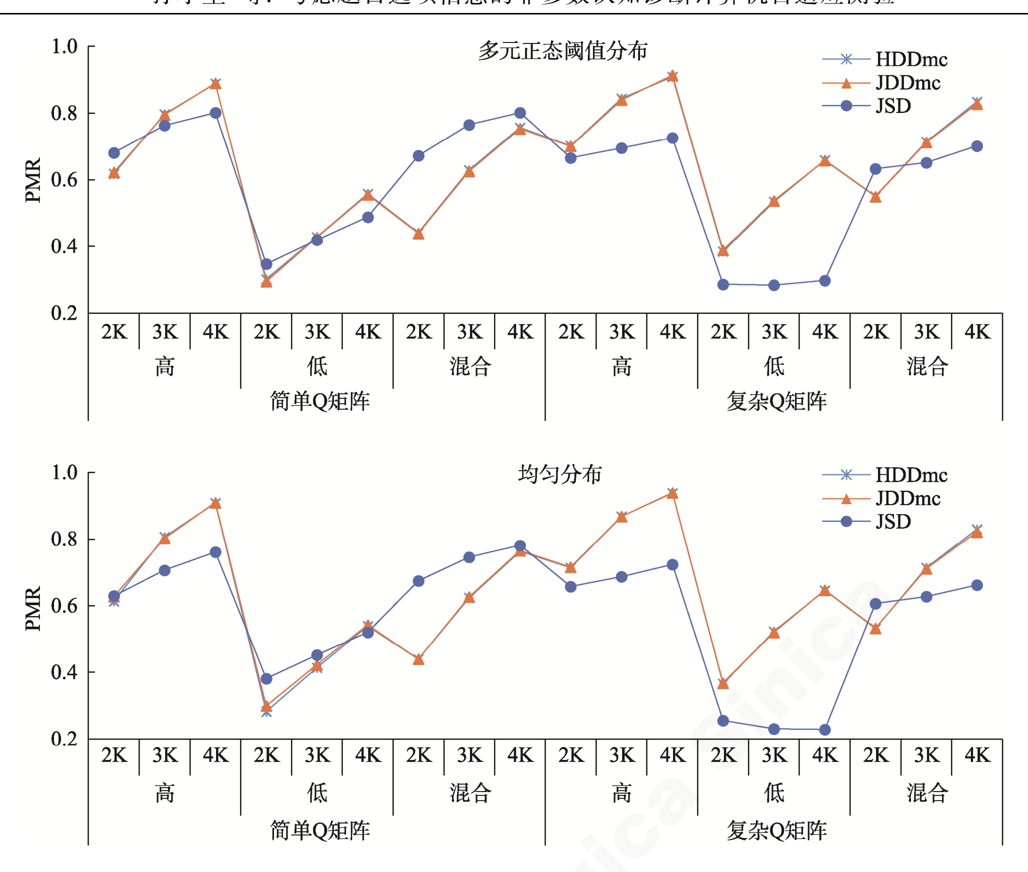


图 2　六个属性下各条件的分类准确性

量较低时, HDDmc 和 JDDmc 的 PMR 明显高于 JSD 方法, 且在 4K 时差异达到最大。与 4 个属性时的结果相同, 题目质量和测验长度对所有选题策略具有正向影响, 题目质量越高、测验长度越长, 则 3 种选题策略的 PMRs 越高。另外, 属性分布形态对选题策略几乎没有影响。

4.4.2　HDDmc 和 JDDmc 的题库使用情况较 JSD 更加均衡

由于 6 个属性下的各选题策略的题库使用情况和 4 个属性时的题库使用情况大体相同, 故不在正文当中呈现, 感兴趣的读者可向作者索要。表 1 呈现了 4 个属性时 3 种选题策略在多元正态阈值分布下的题库使用情况, 总体而言, HDDmc 和 JDDmc 二者在题库使用方面较 JSD 策略更加均衡; HDDmc 和 JDDmc 的题库使用情况基本相同。具体而言, χ^2 方面, HDDmc 和 JDDmc 的 χ^2 分别在 1.167~1.932 和 1.135~1.932 之间, 而 JSD 则在 102.493~199.925 之间, JSD 的整体曝光率远大于 HDDmc 和 JDDmc 二者。测验重叠率(TOR)方面, HDDmc 和 JDDmc 的 TOR 远小于 JSD 的方法, 二者的 TOR 范围均为 0.017~0.035, 而 JSD 的 TOR 范围在 0.229~0.449, 说明 HDDmc 和 JDDmc 在为每个个体选择题目时并没有固定地选择某些共同题目, 而是尽可能地从题库中选择不同的测验题目给个体作答。在曝光不足率(UIR)和过度曝光率(OIR)方面, HDDmc 和 JDDmc 同样表现的要比 JSD 策略更好, 二者的 UIR 和 OIR 均小于 JSD 方法, 特别是 UIR, JSD 策略的 UIR 均在 0.80 以上, 说明使用 JSD 策略时, 题库中存在大量曝光不足的题目。JSD 的 OIR 值虽然比较小(在 0.10 以下), 但 HDDmc 和 JDDmc 的 OIR 均等于 0, 说明这两种非参数选题策略不存在过度曝光的题目, 而 JSD 则存在部分过度曝光的题目。3 种选题策略在均匀分布下的题库使用情况与多元正态阈值分布下的相同, 故不再呈现具体结果: HDDmc 和 JDDmc 的题库使用情况明显好于 JSD, 二者在整体曝光率、测验重叠率、曝光不足率和过度曝光率上的值明显小于 JSD。

5　研究 2: 变长 mcCD-CAT 下两种非参选题策略的性能

5.1　研究目的

探讨 HDDmc 和 JDDmc 在两种新的非参数变长终止规则中的表现情况, 并将其与现有的非参数变长终止规则进行比较。

表 1 四个属性时 3 种策略的题库使用情况(多元正态阈值分布)

题目质量	测验长度	诊断方法	简单 Q 矩阵				复杂 Q 矩阵			
			χ^2	TOR	UIR	OIR	χ^2	TOR	UIR	OIR
高	2K	HDDmc	1.343	0.018	0.663	0.000	1.214	0.017	0.663	0.000
		JDDmc	1.325	0.017	0.667	0.000	1.210	0.017	0.659	0.000
		JSD	148.184	0.324	0.931	0.022	102.493	0.229	0.899	0.022
	3K	HDDmc	1.620	0.026	0.277	0.000	1.432	0.026	0.265	0.000
		JDDmc	1.638	0.026	0.281	0.000	1.431	0.026	0.259	0.000
		JSD	171.692	0.381	0.905	0.045	121.801	0.277	0.869	0.043
	4K	HDDmc	1.932	0.035	0.096	0.000	1.659	0.035	0.078	0.000
		JDDmc	1.932	0.035	0.093	0.000	1.660	0.035	0.077	0.000
		JSD	187.661	0.423	0.875	0.059	133.228	0.310	0.839	0.060
低	2K	HDDmc	1.236	0.017	0.669	0.000	1.167	0.017	0.666	0.000
		JDDmc	1.216	0.017	0.660	0.000	1.135	0.017	0.668	0.000
		JSD	187.023	0.405	0.934	0.023	139.115	0.305	0.909	0.023
	3K	HDDmc	1.392	0.026	0.256	0.000	1.273	0.026	0.242	0.000
		JDDmc	1.413	0.026	0.256	0.000	1.250	0.026	0.239	0.000
		JSD	194.582	0.429	0.902	0.039	152.600	0.342	0.873	0.043
	4K	HDDmc	1.581	0.035	0.064	0.000	1.418	0.034	0.064	0.000
		JDDmc	1.581	0.035	0.065	0.000	1.425	0.034	0.065	0.000
		JSD	199.925	0.449	0.872	0.050	165.428	0.377	0.840	0.053
混合	2K	HDDmc	1.260	0.017	0.664	0.000	1.169	0.017	0.665	0.000
		JDDmc	1.292	0.017	0.663	0.000	1.143	0.017	0.662	0.000
		JSD	173.083	0.376	0.938	0.028	142.062	0.311	0.918	0.021
	3K	HDDmc	1.509	0.026	0.264	0.000	1.343	0.026	0.249	0.000
		JDDmc	1.537	0.026	0.268	0.000	1.378	0.026	0.255	0.000
		JSD	191.497	0.423	0.901	0.035	173.288	0.385	0.898	0.037
	4K	HDDmc	1.737	0.035	0.082	0.000	1.546	0.035	0.072	0.000
		JDDmc	1.763	0.035	0.083	0.000	1.589	0.035	0.071	0.000
		JSD	192.640	0.434	0.868	0.051	181.812	0.411	0.867	0.052

5.2 研究设计

5.2.1 自变量

研究 2 的自变量个数有 6 个，其中属性个数、Q 矩阵结构、题目质量和属性分布形态 4 个自变量的设定与研究 1 相同。剩余两个自变量为终止规则和选题策略，终止规则有 4 个水平，分别为张淑君(2019)提出的 D1 和 D3 规则，以及本研究中的 MR 和 DR 规则。选题策略方面，由于参数终止规则无法与非参数选题策略匹配，故未考虑参数选题策略(JSD)，而重点关注 HDDmc 和 JDDmc 二者在不同终止规则下的表现。

5.2.2 控制变量

研究 2 中的控制变量和题库与研究 1 相同，测试人数的真实 AMPs 重新生成。由于 DR 规则需预先设定 CR 值，基于预实验的结果，将 4 个和 6 个属性下的 CR 值分别设为 1.3 和 1.25 时，HDDmc 和 JDDmc 可获得较好的结果，故本研究使用这两个值。测验长度的上限设置为 30 题(郭磊 等，2015，2016)。此外，为防止测验未测量所有属性而导致提前终止的情况，研究使用 Xu 等(2016)的初始题目选择程序以保证每个个体在每个属性上均提供了相应的作答信息。

研究总共有 2 (属性个数) × 2 (Q 矩阵结构) × 3 (题目质量) × 2 (属性分布形态) × 4 (终止规则) × 2 (选题策略) = 192 种实验条件，其中终止规则和选题策略为被试内变量，其它则为被试间变量。各条件重复 30 次。所有程序用 R 软件实现。

5.3 评价指标

研究的评价指标同样分为准确性指标和题目使用情况，其中准确性指标为 PMR，而题目使用情

况的指标则为平均测验长度(M)、最小测验长度(Min)、最大测验长度(Max)、UIR 和 OIR (郭磊 等, 2016; 孙小坚 等, 2021)。

5.4 研究结果

表 2 和表 3 分别呈现了多元正态阈值分布下, HDDmc 和 JDDmc 在 4 个和 6 个属性的表现情况, 整体而言, HDDmc 和 JDDmc 在 MR 和 DR 两种终止规则下的分类准确性较 D1 和 D3 高, 但测验长度更长; 同时二者在曝光不足率上的表现优于 D1 和 D3。下面分别对两个表格进行阐述。表 2 呈现了 HDDmc 和 JDDmc 在 4 个属性和多元正态阈值分布条件下的分类准确性(PMR)以及题库使用情况。MR 和 DR 规则下, HDDmc 和 JDDmc 的 PMRs 范围为 0.441~0.775 (M =0.659); 二者在 D1 和 D3 规则下的 PMRs 则为 0.288~0.703 (M = 0.475)。测验长度的使用方面, HDDmc 和 JDDmc 在 MR 和 DR 规则下的平均测验长度、最小测验长度以及最大测验长度三者均要大于 D1 和 D3 规则下的使用情况。HDDmc 和 JDDmc 在 D1 和 D3 上的平均、最小和最大题目长度的范围分别为 5.289~8.319、5.0~7.0 和 8.667~14.90; 而二者在 MR 和 DR 规则下的平均、最小和最大题目长度则分别为 9.274~20.838、5.0~7.0 和 25.033~30.0。题目曝光率方面, HDDmc 和 JDDmc 在 MR 和 DR 规则下曝光不足率(UIR)明显小于二者在 D1 和 D3 规则下的 UIR, MR 和 DR 规则下的 UIR 为 0.003~0.661 (M = 0.345), 而 D1 和 D3 规则下的 UIR 则为 0.608~0.849 (M = 0.737), 说明 HDDmc 和 JDDmc 在 D1 和 D3 规则下存在大量曝光不足的题目, 而 MR 和 DR 规则下曝光不足的题目则较少; 此外, 所有终止规则下的过度曝光率(OIR)均为 0, 说明两种非参数选题策略在不同终止规则下均不存在过度曝光的题目。均匀分布下的分类结果和题库使用情况与多元正态阈值分布下的相同, 将不再呈现。

表 2　四个属性时两种非参方法的分类结果及题库使用情况(多元正态阈值分布)

题目质量	终止规则	诊断方法	简单 Q 矩阵 [a]					复杂 Q 矩阵 [a]				
			M	Min	Max	UIR	PMR	M	Min	Max	UIR	PMR
高	MR	HDDmc	9.274	7	25.033	0.520	0.712	9.369	7	26.333	0.661	0.775
		JDDmc	9.300	7	25.367	0.511	0.710	9.402	7	25.600	0.659	0.768
	DR	HDDmc	13.785	5	30.000	0.134	0.724	11.767	5	30.000	0.473	0.738
		JDDmc	14.876	5	30.000	0.086	0.745	12.853	5	30.000	0.363	0.752
	D1	HDDmc	5.308	5	8.733	0.849	0.496	5.289	5	9.400	0.751	0.514
		JDDmc	5.303	5	8.667	0.846	0.490	5.293	5	9.200	0.752	0.508
	D3	HDDmc	7.939	7	12.833	0.650	0.648	7.914	7	13.433	0.728	0.703
		JDDmc	7.964	7	12.900	0.651	0.651	7.920	7	13.600	0.726	0.702
低	MR	HDDmc	10.204	7	28.667	0.414	0.450	10.482	7	29.167	0.577	0.509
		JDDmc	10.199	7	28.367	0.415	0.441	10.460	7	29.133	0.582	0.514
	DR	HDDmc	19.536	5	30.000	0.009	0.629	17.237	5	30.000	0.097	0.648
		JDDmc	20.838	5	30.000	0.003	0.641	18.545	5	30.000	0.068	0.663
	D1	HDDmc	5.431	5	9.333	0.839	0.288	5.430	5	10.300	0.750	0.303
		JDDmc	5.423	5	9.333	0.841	0.293	5.418	5	10.367	0.751	0.310
	D3	HDDmc	8.308	7	13.833	0.612	0.396	8.319	7	14.900	0.716	0.445
		JDDmc	8.315	7	13.800	0.608	0.397	8.303	7	14.733	0.719	0.434
混合	MR	HDDmc	9.762	7	26.400	0.463	0.591	9.961	7	27.233	0.620	0.666
		JDDmc	9.765	7	25.867	0.466	0.595	9.915	7	27.733	0.619	0.665
	DR	HDDmc	16.321	5	30.000	0.042	0.720	13.902	5	30.000	0.277	0.711
		JDDmc	17.570	5	30.000	0.027	0.729	15.141	5	30.000	0.192	0.724
	D1	HDDmc	5.368	5	8.833	0.839	0.379	5.364	5	10.033	0.750	0.416
		JDDmc	5.368	5	9.000	0.845	0.391	5.352	5	9.700	0.750	0.418
	D3	HDDmc	8.138	7	13.200	0.633	0.521	8.135	7	14.467	0.722	0.585
		JDDmc	8.131	7	13.233	0.629	0.530	8.125	7	13.867	0.723	0.589

注: MR 表示基于限制性 MHRM 算法的终止规则, DR 表示基于距离比的终止规则; [a] 表示 OIR 均为 0。

表 3　六个属性时两种非参方法的分类结果及题库使用情况(多元正态阈值分布)

题目质量	终止规则	诊断方法	简单 Q 矩阵 [a]					复杂 Q 矩阵 [b]				
			M	Min	Max	UIR	PMR	*M*	Min	Max	UIR	PMR
高	MR	HDDmc	11.814	9	29.067	0.381	0.536	12.226	9	29.000	0.381	0.536
		JDDmc	11.845	9	28.967	0.380	0.529	12.218	9	29.400	0.380	0.529
	DR	HDDmc	12.628	7	30.000	0.351	0.485	11.571	7	30.000	0.351	0.485
		JDDmc	16.169	7	30.000	0.113	0.603	14.532	7	30.000	0.113	0.603
	D1	HDDmc	7.320	7	11.167	0.574	0.319	7.295	7	11.900	0.574	0.319
		JDDmc	7.317	7	11.300	0.576	0.316	7.287	7	11.700	0.576	0.316
	D3	HDDmc	9.988	9	15.767	0.493	0.440	10.015	9	17.533	0.493	0.440
		JDDmc	9.984	9	15.600	0.496	0.442	9.992	9	16.533	0.496	0.442
低	MR	HDDmc	12.946	9	29.900	0.307	0.273	13.599	9	30.000	0.307	0.273
		JDDmc	12.934	9	29.867	0.308	0.274	13.463	9	30.000	0.308	0.274
	DR	HDDmc	16.994	7	30.000	0.098	0.353	15.286	7	30.000	0.098	0.353
		JDDmc	21.679	7	30.000	0.014	0.457	19.653	7	30.000	0.014	0.457
	D1	HDDmc	7.428	7	12.000	0.572	0.141	7.434	7	12.900	0.572	0.141
		JDDmc	7.436	7	11.800	0.569	0.146	7.427	7	12.967	0.569	0.146
	D3	HDDmc	10.332	9	16.533	0.484	0.210	10.396	9	17.833	0.484	0.210
		JDDmc	10.352	9	16.867	0.482	0.213	10.404	9	18.367	0.482	0.213
混合	MR	HDDmc	12.468	9	29.633	0.335	0.400	12.834	9	29.833	0.335	0.400
		JDDmc	12.438	9	29.633	0.337	0.389	12.954	9	29.967	0.337	0.389
	DR	HDDmc	14.683	7	30.000	0.204	0.423	12.770	7	30.000	0.204	0.423
		JDDmc	19.093	7	30.000	0.036	0.559	16.560	7	30.000	0.036	0.559
	D1	HDDmc	7.385	7	11.533	0.570	0.212	7.376	7	12.733	0.570	0.212
		JDDmc	7.390	7	11.533	0.573	0.210	7.367	7	12.933	0.573	0.210
	D3	HDDmc	10.202	9	16.233	0.486	0.314	10.198	9	17.533	0.486	0.314
		JDDmc	10.179	9	16.500	0.488	0.309	10.234	9	17.800	0.488	0.309

注：[a] 表示简单 *Q* 矩阵结构下的 OIR 均为 0; [b] 表示复杂 Q 矩阵结构下的 OIR 均为 0.008。

表 3 呈现了 HDDmc 和 JDDmc 在 6 个属性和多元正态阈值分布条件下的分类准确性(PMR)以及题库使用情况。MR 和 DR 规则下，HDDmc 和 JDDmc 的 PMRs 范围为 0.273~0.639 (M = 0.471); 二者在 D1 和 D3 规则下的 PMRs 则为 0.141~0.511 (M = 0.296)。测验长度的使用方面，HDDmc 和 JDDmc 在 D1 和 D3 上的平均、最小和最大题目长度的范围分别为 7.287~10.404、7.0~9.0 和 11.167~18.367; 而二者在 MR 和 DR 规则下的平均、最小和最大题目长度则分别为 11.571~21.679、7.0~9.0 和 28.967~30.0。题目曝光率方面, HDDmc 和 JDDmc 在 MR 和 DR 规则下的 UIR 为 0.036~0.811 (M = 0.412), 而 D1 和 D3 规则下的 UIR 则为 0.482~0.902 (M = 0.711)。此外，所有终止规则下的过度曝光率(OIR)均非常小，说明两种非参数选题策略在不同终止规则下均难以产生过度曝光的题目。均匀分布下的分类结果和题库使用情况与多元正态阈值分布下的相同，故不再呈现。

6　讨论与结论

6.1　研究讨论

当前大部分 CD-CAT 的研究常忽略干扰项的诊断信息，造成资源的浪费，对此 Yigit 等(2019)基于 MC-DINA 模型提出了综合使用题目所有选项信息的参数选题策略，并取得理想结果。但参数方法面临计算复杂、前提假设严苛以及需较大样本量等不足(郭磊 等, 2018; 康春花 等, 2015; Chiu et al., 2018)。基于此，本研究提出了两种适用于 mcCD-CAT 的非参数选题策略(HDDmc 和 JDDmc)，并且还提出两种变长 CD-CAT 情境下的终止规则。通过两个模拟研究系统地探讨了二者在 mcCD-CAT 中的表现情况。结果发现，定长实验条件下，非参数选题策略 HDDmc 和 JDDmc 可以获得较参数选题策略更加准确的分类结果，并且其题库使用情况明

显好于参数选题策略。

6.1.1 控制简单 Q 矩阵和混合题目质量条件下，属性个数对非参数选题策略有消极影响

模拟研究 1 的结果显示，4 个属性时，HDDmc 和 JDDmc 在简单 Q 矩阵和混合题目质量下的分类准确性整体要优于 JSD 方法，但在 6 个属性时，JSD 的分类准确性则高于 HDDmc 和 JDDmc，特别是测验长度为 2K 和 3K 时。该结果产生的可能原因是 4 个属性时，HDDmc 和 JDDmc 倾向于选择特定的题目集，而 6 个属性时，二者所选择的题目集范围更广泛。当候选题目集范围较为广泛时，由于是从题目集中随机选择一个题目，故导致非参数选题策略可能无法获得最佳的测验题目，从而产生较低的分类准确性；而参数选题策略则可以计算各个题目的 JSD，再确定性地从题目集中选择具有最大 JSD 值的题目。这也许可以从题库使用情况对其进行论证：HDDmc 和 JDDmc 在 4 和 6 个属性下的整体曝光率、测验重叠率和过度曝光率三个方面的差异比较小，但曝光不足率方面，二者在 4 和 6 个属性上的差异则比较大，说明 HDDmc 和 JDDmc 在 4 个属性下存在大量曝光不足的题目，这一定程度上反向说明该条件下 HDDmc 和 JDDmc 倾向于选择特定的某些题目集。

6.1.2 MR 和 DR 规则在平衡准确性和题库使用间的表现较 D1 和 D3 规则稍差

模拟研究 2 的结果显示，研究提出的两种新的非参数终止规则可以获得较 D1 和 D3 更高的分类准确性，但其代价则是需要更多的测验题目，特别是 DR 规则，其所需的题目数明显多于其它 3 种规则，该规则下的平均测验长度均在 10.0 以上。当然，这也跟研究的设定有关，MR 规则下，个体需连续获得 4 个完全一致性的 AMP 值时测验方能结束，而 DR 规则下，第二小和最小的距离之间的比值需在 1.3 或 1.25 时，测验才能结束，这些设定相对于 D1 和 D3 而言，更加严苛，因而其需要更多的测验题目，进而导致更高的分类准确性。这是 CAT 情境中一直面临的利益权衡问题(陈平 等, 2011; 郭磊 等, 2015; 毛秀珍, 辛涛, 2013; 孙小坚 等, 2021)。事实上，MR 和 DR 规则下分类准确性的高低和题库使用情况之间的利益权衡可通过研究设置给予实现，当研究目的在于尽可能获得准确分类结果时，可增加 MR 规则下连续一致性 AMP 值的次数和增大 DR 规则中的 CR 值；反之，则可以适当减少。

6.1.3 Q 矩阵复杂程度对分类准确性有正向影响

两个模拟研究的结果还显示，相对于简单 Q 矩阵，3 种选题方法在复杂 Q 矩阵下的分类准确性更高。其原因可能在于简单 Q 矩阵情境下，题库中大部分题目只测量了一个属性(本研究中 4 和 6 个属性时各有 317 和 252 个题目)，这些题目的干扰项没有提供任何额外信息，因此简单 Q 矩阵中的题目提供的选项信息有限。而复杂 Q 矩阵下，只测量一个属性的题目比例则比较少(本研究中 4 和 6 个属性下分别仅有 26.25%和 9.79%的比例)，剩余题目的干扰项均能提供诊断信息，因此在复杂 Q 矩阵下可得到更高的分类准确性。

6.1.4 HDDmc 和 JDDmc 不依赖于预测试的样本量

基于两个模拟研究的分析过程可以发现，在正式测试之前，需进行预测试以获得题目参数的估计值，从而为后续的正式测试提供题目参数信息。而前人研究发现，预测试的样本量会影响参数选题策略的估计准确性，预测试样本量越大，则参数选题策略的估计准确性也越高(Huang, 2018; Sun et al., 2020)。其原因在于样本量较小时，参数估计的误差将比较大，而参数选题策略直接将误差较大的题目参数估计值作为正式测试中的真值，从而影响个体 AMP 的估计准确性。如此，可以预期，较小的预测试样本量将影响 JSD 的分类结果。反观 HDDmc 和 JDDmc，二者不需要进行预测试，因而预测试的样本量大小不会对其产生影响，该结果与以往关于非参数诊断方法的研究结果相同(如：康春花 等, 2019; 罗照盛 等, 2015)。

6.1.5 研究不足与展望

本研究丰富了关于 mcCD-CAT 的研究。当然，后续研究还可从以下几个方面进行深入探究：(1) *Q* 矩阵方面，MC-DINA 模型要求干扰项的 *q* 向量必须是正确选项的子集，但实际的测验编制过程中，干扰项的 q 向量不属于正确选项的子集同样有可能发生(郭磊, 周文杰, 2021)，因此后续研究可对此进行探讨。(2)研究只考虑了个体在各选项上的作答情况，其他信息如作答时间等变量同样可以提供额外的诊断信息，后续研究可尝试将时间信息给予考虑。(3)研究为模拟研究，各方面可以进行严格控制，而实际测验情境将会更加复杂，因此，非参数方法在实证研究中的效果如何需要进一步验证。

6.2 研究结论

基于两个模拟研究的结果，研究得到以下结

论：(1)两种非参数选题策略均适用于 mcCD-CAT 情境，二者均获得较高的分类准确性，因此，使用者可以任选其一；(2)两种非参数方法具有较为均匀的题库使用情况，一定程度上保证了题库的安全性；(3)两种非参数终止规则适用于变长 mcCD-CAT 情境，可依据测验目的灵活地平衡准确性和题库使用情况；当测验追求精度时，MR 规则的连续相等次数可设置为 5 次及以上，而 DR 规则下的 CR 值则可以设置为 1.5 及以上；反之，则可以降低 MR 规则中的次数和 DR 规则中的 CR 值。

参 考 文 献

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*(3), 403–425.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*(1), 1–20.

Chang, Y.-P., Chiu, C.-Y., & Tsai, R.-C. (2019). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*, *43*(7), 543–561.

Chen, P., Li, Z., & Xin, T. (2011). A note on the uniformity of item bank usage in cognitive diagnostic computerized adaptive testing. *Studies of Psychology and Behavior*, *9*(2), 125–132.

[陈平, 李珍, 辛涛. (2011). 认知诊断计算机化自适应测验的题库使用均匀性初探. *心理与行为研究*, *9*(2), 125–132.]

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619–632.

Chiu, C. Y., & Chang, Y. P. (2021). Advances in CD-CAT: The general nonparametric item selection method. *Psychometrika, 86*(4), 1039–1057.

Chiu, C. Y., Douglas, J. A., & Li, X. D. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633–665.

Chiu, C. Y., Sun, Y., & Bian, Y. H. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika, 83*(2), 355–375.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163–183.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

Gao, Y., Zhai, X., Cui, Y., Xin, T., & Bulut, O. (2021). Re-validating a learning progression of buoyancy for middle school students: A longitudinal study. *Research in Science Education*. Advance online publication. https://doi.org/10.1007/s11165-021-10021-x.

Guo, L., Yang, J., & Song, N. Q. (2018). Application of spectral clustering algorithm under various attribute hierarchical structures for cognitive diagnostic assessment. *Journal of Psychological Science, 41*(3), 735–742.

[郭磊, 杨静, 宋乃庆. (2018). 谱聚类算法在不同属性层级结构诊断评估中的应用. *心理科学, 41*(3), 735–742.]

Guo, L., Zheng, C., Bian, Y. (2015). Exposure control methods and termination rules in variable-length cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica*, *47*(1), 129–140.

[郭磊, 郑蝉金, 边玉芳. (2015). 变长 CD-CAT 中的曝光控制与终止规则. *心理学报, 47*(1), 129–140.]

Guo, L., Zheng, C., Bian, Y., Song, N., & Xia, L. (2016). New item selection methods in cognitive diagnostic computerized adaptive testing: Combining item discrimination indices. *Acta Psychologica Sinica*, *48*(7), 903–914.

[郭磊, 郑蝉金, 边玉芳, 宋乃庆, 夏凌翔. (2016). 认知诊断计算机化自适应测验中新的选题策略：结合项目区分度指标. *心理学报, 48*(7), 903–914.]

Guo, L., & Zhou, W. (2021). Nonparametric methods for cognitive diagnosis to multiple-choice test items. *Acta Psychologica Sinica, 53*(9), 1032–1043.

[郭磊, 周文杰. (2021). 基于选项层面的认知诊断非参数方法. *心理学报, 53*(9), 1032–1043.]

He, M. (2021). *Research on nonparametric cognitive diagnosis method and item selection strategy of nonparametric CD-CAT* (Unpublished master's thesis). Sichuan Normal University, Chengdu, China.

[何明霜. (2021). *非参数认知诊断方法与非参数 CD-CAT 选题策略研究* (硕士毕业论文). 四川师范大学, 成都.]

Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*(7), 563–582.

Huang, H.-Y. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *Journal of Classification, 35*(3), 437–465.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist, 11*(2), 37–50.

Kang, C. H., Ren, P., & Zeng, P. F. (2015). Nonparametric cognitive diagnosis: A cluster diagnostic method based on grade response items. *Acta Psychologica Sinica, 47*(8), 1077–1088.

[康春花, 任平, 曾平飞. (2015). 非参数认知诊断方法：多级评分的聚类分析. *心理学报, 47*(8), 1077–1088.]

Kang, C. H., Yang, Y. K., & Zeng, P. F. (2019). Approach to cognitive diagnosis: The Manhattan distance discriminating method. *Journal of Psychological Science, 42*(2), 455–462.

[康春花, 杨亚坤, 曾平飞. (2019). 一种混合计分的非参数认知诊断方法：曼哈顿距离判别法. *心理科学, 42*(2), 455–462.]

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *39*(3), 167–188.

Kosub, S. (2019). A note on the triangle inequality for the Jaccard distance. *Pattern Recognition Letters*, *120*, 36–38.

Liu, C., & Cheng, Y. (2018). An application of the support vector machine for attribute-by-attribute classification in cognitive diagnosis. *Applied Psychological Measurement, 42*(1), 58–72.

Liu, C.-W., Andersson, B., & Skrondal, A. (2020). A constrained Metropolis–Hastings Robbins–Monro algorithm for Q matrix estimation in DINA models. *Psychometrika, 85*(2), 322–357.

Liu, H. Y., You, X. F., Wang, W.-Y., Ding, S.-L., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30*(2), 152–172.

Liu, T. (2016). *Using distractor information in computerized adaptive testing* (Unpublished doctoral dissertation). Beijing Normal University, Beijing.

[刘拓. (2016). *干扰项信息在计算机化自适应测验中的利用* (博士学位论文). 北京师范大学, 北京.]

Luo, Z., Li, Y., Yu, X., Gao, C., & Peng, Y. (2015). A simple

cognitive diagnosis method based on *Q*-matrix theory. *Acta Psychologica Sinica*, *47*(2), 264–272.
[罗照盛, 李喻骏, 喻晓锋, 高椿雷, 彭亚风. (2015). 一种基于 Q 矩阵理论朴素的认知诊断方法. *心理学报*, *47*(2), 264–272.]
Mao, X. Z., & Xin, T. (2013). A comparison of item selection methods for controlling exposure rate in cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica*, *45*(6), 694–703.
[毛秀珍, 辛涛. (2013). 认知诊断 CAT 中项目曝光控制方法的比较. *心理学报*, *45*(6), 694–703.]
Ozaki, K. (2015). DINA models for multiple-choice items with few parameters: Considering incorrect answers. *Applied Psychological Measurement*, *39*(6), 431–447.
Sun, X., Andersson, B., & Xin, T. (2021). A new method to balance measurement accuracy and attribute coverage in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *45*(7–8), 463–476.
Sun, X., Liu, Y., Xin, T., & Song, N. (2020). The impact of item calibration error on variable-length cognitive diagnostic computerized adaptive testing. *Frontiers in Psychology*, *141*(11), Article e575141. https://10.3389/fpsyg.2020.575141
Sun, X., Mao, X., Song, N., & Xin, T. (2021). New methods for item exposure control in cognitive diagnostic computerized adaptive testing. *Journal of Psychological Science*, *44*(1), 205–213.
[孙小坚, 毛秀珍, 宋乃庆, 辛涛. (2021). 定长 CD-CAT 中两种新的题目曝光控制方法. *心理科学*, *44*(1), 205–213.]
Sun, X., Wang, Y., Zhang, S., & Xin, T. (2019). New methods to balance attribute coverage for cognitive diagnostic computerized adaptive testing. *Journal of Psychological Science*, *42*(5), 1236-1244.
[孙小坚, 王钰彤, 张世夷, 辛涛. (2019). 认知诊断计算机自适应测验中平衡属性收敛的新方法. *心理科学*, *42*(5), 1236–1244.]
Wang, W. Y., Ding, S. L., Song, L. H., Kuang, Z., & Cao, H. Y. (2016). Application of neural networks and support vector machines to cognitive diagnosis. *Journal of Psychological Science*, *39*(4), 777–782.
[汪文义, 丁树良, 宋丽红, 邝铮, 曹慧媛. (2016). 神经网络和支持向量机在认知诊断中的应用. *心理科学*, *39*(4), 777–782.]
Xin, T., Le, M., & Guo, Y., & Jiang, Y. (2015). The approach to establishing achievement standard: The learning progressions based on cognition diagnostic. *Journal of Educational Studies*, *5*, 72–79.
[辛涛, 乐美玲, 郭艳芳, 姜宇. (2015). 学业质量标准的建立途径：基于认知诊断的学习进阶方法. *教育学报*, *5*, 72–79.]
Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 291–315.
Xu, X. L., Chang, H.-H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.
Yamaguchi, K. (2020). Variational Bayesian inference for the multiple-choice DINA model. *Behaviormetrika*, *47*(1), 159–187.
Yang, J., Chang, H.-H., Tao, J., & Shi, N. (2020). Stratified item selection methods in cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *44*(5), 346–361.
Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement*, *43*(5), 388–401.
Zhang, S. (2019). *Applying npCD-CAT based on MDD to the field of number and algebra* (Unpublished master's thesis). Zhejiang Normal University, Jinhua, China.
[张淑君. (2019). *基于 MDD 的 npCD-CAT 研究及其在数与代数领域的应用* (硕士毕业论文). 浙江师范大学, 金华.]
Zheng, C. J., & Chang, H.-H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *40*(8), 608–624.

Nonparametric cognitive diagnostic computerized adaptive testing using multiple-choice option information

SUN Xiaojian[1,2,3], GUO Lei[3,4]

([1] *School of Mathematics and Statistics, Southwest University, Chongqing* 400715, *China*)
([2] *Basic Education Research Centre, Southwest University, Chongqing* 400715, *China*)
([3] *Southwest University Branch, Collaborative Innovation Center of Assessment for Basic Education Quality, Chongqing* 400715, *China*)
([4] *Faculty of Psychology, Southwest University, Chongqing* 400715, *China*)

Abstract

Most existing cognitive diagnostic computerized adaptive testing (CD-CAT) item selection methods ignore the diagnostic information that distractors provide for multiple-choice (MC) items. Consequently, some useful information is missed and resources are wasted. To overcome this, researchers proposed the Jensen–Shannon divergence (JSD) strategy to select items with the MC-DINA model. However, the JSD strategy needs large samples to obtain reliable estimates of the item parameters before the formal test, and this could compromise the

items in the bank. By contrast, the nonparametric method does not require any parameter calibration before the formal test and can be used in small educational programs.

The current study proposes two nonparametric item selection methods (i.e., HDDmc and JDDmc) for CD-CAT with MC items as well as two termination rules (i.e., MR and DR) for variable-length CD-CAT with MC items. Two simulation studies were conducted to examine the performance of these nonparametric item selection methods and termination rules.

The first study examined the performance of the HDDmc and JDDmc with fixed-length CD-CAT. In this study, six factors were manipulated: the number of attributes (K = 4 vs. 6), the structure of the Q-matrix (simple vs. complex), the quality of the item bank (high vs. low vs. mixed), the distribution of the attribute profile (multivariate normal threshold model vs. discrete uniform distribution), the test length (two vs. three vs. four times of K), and the item selection methods (HDDmc vs. JDDmc vs. JSD). Of these, item selection method was the within-group variable, and the rest were between-group variables. The results showed that: (1) the HDDmc and JDDmc produced higher attribute pattern matched ratios (PMRs) than the JSD method for most conditions; (2) the HDDmc and JDDmc produced similar PMRs for all conditions; (3) the HDDmc and JDDmc produced more even distributions of item exposure than the JSD method.

The second simulation study investigated the performance of the MR and DR with variable-length CD-CAT. Six factors were also manipulated in this study: the settings for the number of attributes, the structure of the Q-matrix, the quality of the item bank, and the distribution of the attribute profile were the same as in the first study; the other two factors were termination rules (MR, DR, D1, and D3) and item selection methods (HDDmc and JDDmc). Again, the first four were between-group variables, while termination rules and item selection methods were within-group variables. The results showed that: (1) the HDDmc and JDDmc yielded higher PMRs for MR and DR rules than for the D1 and D3 rules; (2) the HDDmc and JDDmc yielded longer test lengths for MR and DR rules than for the D1 and D3 rules, especially for the JDD rule.

In sum, both nonparametric item selection methods and the two new termination rules proved appropriate for CD-CAT with MC items, which means they can be used to balance the trade-off between measurement accuracy and item exposure rate.

Key words cognitive diagnostic computerized adaptive testing, multiple-choice items, nonparametric item selection method, termination rule